# Nonparametric Applications of Bayesian Inference

TECHNICAL WORKING PAPER SERIES

# NONPARAMETRIC APPLICATIONS OF BAYESIAN INFERENCE

Gary Chamberlain
Guido W. Imbens

Technical Working Paper 200

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 1996

The authors thank David Cox, Jinyong Hahn, and Neil Shephard for helpful comments, and thank Alan Krueger and Bruce Meyer for making their data available to us. The National Science Foundation provided financial support. This paper is part of NBER's research program in Labor Studies. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Technical Working Paper 200
August 1996

# NONPARAMETRIC APPLICATIONS OF BAYESIAN INFERENCE

## ABSTRACT

The paper evaluates the usefulness of a nonparametric approach to Bayesian inference by presenting two applications. The approach is due to Ferguson (1973, 1974) and Rubin (1981). Our first application considers an educational choice problem. We focus on obtaining a predictive distribution for earnings corresponding to various levels of schooling. This predictive distribution incorporates the parameter uncertainty, so that it is relevant for decision making under uncertainty in the expected utility framework of microeconomics. The second application is to quantile regression. Our point here is to examine the potential of the nonparametric framework to provide inferences without making asymptotic approximations. Unlike in the first application, the standard asymptotic normal approximation turns out to not be a good guide. We also consider a comparison with a bootstrap approach.

Gary Chamberlain
Department of Economics
Littauer 123
Harvard University
Cambridge, MA 02138
and NBER
CHAMBERL@ECGC.HARVARD.EDU

Guido W. Imbens
Department of Economics
Littauer 117
Harvard University
Cambridge, MA 02138
and NBER
guido_imbens@harvard.edu

# NONPARAMETRIC APPLICATIONS OF BAYESIAN INFERENCE[1]

## 1. INTRODUCTION

The paper evaluates the usefulness of a nonparametric approach to Bayesian inference by presenting two applications. The approach is due to Ferguson (1973, 1974) and Rubin (1981). It stays within the framework of evaluating expected utility with respect to a posterior distribution, but it does not use a tightly parameterized likelihood function based, for example, on a normal distribution. At the same time, it avoids pitfalls arising from using high-dimensional parameter spaces with flat or other conventional prior distributions.

Our first application considers an educational choice problem. We focus on obtaining a predictive distribution for earnings corresponding to various levels of schooling. This predictive distribution incorporates the parameter uncertainty, so that it is relevant for decision making under uncertainty in the expected utility framework of microeconomics. Specifically, we look at an individual's decision on the level of schooling when the individual is uncertain about the return to schooling.

The second application is to quantile regression. Our point here is to examine the potential of the nonparametric framework to provide inferences without making asymptotic approximations. Unlike in the first application, the standard asymptotic normal approximation turns out to not be a good guide. We also consider a comparison with a bootstrap approach.

## 2. THEORY

Our aim is to present a concise review of the basic theory that is sufficient to follow the applications. For more details, see Ferguson (1973, 1974), Rubin (1981), and Chamberlain and Imbens (1995). There is a family of probability distributions $\{P_\theta : \theta \in \Theta\}$, and we observe $\{Z_i\}_{i=1}^n$, where the random variables $Z_i$ are independently and identically

distributed according to $P_\theta$ for some unknown value of $\theta$ in the parameter space $\Theta$. To simplify notation, let $Z$ denote a random variable that is distributed according to $P_\theta$. We assume that the distributions $P_\theta$ have a common, finite support:

$$P_\theta(Z = a_j) = \theta_j \qquad (j = 1, \ldots, J),$$

where $\theta_j$ denotes the $j^{\text{th}}$ component of $\theta$ and we take $\Theta$ to be the unit simplex in $\mathcal{R}^J$. Since $J$ can be arbitrarily large, and our data are measured with finite precision, the finite support assumption need not be restrictive.

We limit ourselves to prior distributions in the Dirichlet family with density

$$p(\theta) \propto \prod_{j=1}^{J} \theta_j^{b_j - 1} \quad \text{for} \quad \theta \in \Theta \qquad (b_j > 0)$$

which, with $J$ free parameters, is arguably not very restrictive. Let $d = \{z_i\}_{i=1}^n$ denote the data, that is, the observed values of the $Z_i$. The posterior density is proportional to the product of the prior density and the likelihood function:

$$p_n(\theta \mid d) \propto \prod_{j=1}^{J} \theta_j^{n_j + b_j - 1}, \tag{1}$$

where $n_j = \sum_{i=1}^n 1(z_i = a_j)$ is the number of sample observations equal to $a_j$. This is the standard result that the posterior distribution is also Dirichlet.

The Dirichlet distribution has a very convenient representation.[2] Let $\{U_j\}_{j=1}^J$ be independent random variables with $U_j$ distributed according to a gamma distribution with shape parameter $b_j$ and scale parameter 1: $U_j \sim \mathcal{G}(b_j, 1)$. Then $(U_1, \ldots, U_J)/\sum_{j=1}^J U_j$ is distributed Dirichlet with parameters $\{b_j\}_{j=1}^J$. This representation is correct even if some (but not all) of the $b_j$ equal 0, if we interpret the $\mathcal{G}(0, 1)$ distribution to assign unit probability to 0. If $b_j > 0$ for $j \leq K$ and $b_j = 0$ for $j > K$, the Dirichlet distribution for $\theta = (\theta_1, \ldots, \theta_J)$ has $\theta_j = 0$ for $j > K$ with probability one and $(\theta_1, \ldots, \theta_K)$ has a Dirichlet distribution with (positive) parameters $b_1, \ldots, b_K$.

Suppose that we are interested in some function of $\theta$: $\beta = g(\theta)$, where the function $g(\cdot)$ may depend upon the points of support $\{a_j\}_{j=1}^J$. The representation for the Dirichlet posterior distribution can be used to simulate the posterior distribution of $\beta$ by taking independent draws from a gamma distribution. Given $L$ independent draws $\{\theta^{(l)}\}_{l=1}^L$ from $p_n(\theta \mid d)$, we can set $\beta^{(l)} = g(\theta^{(l)})$ to obtain independent draws from the posterior distribution of $\beta$. We define $g(\cdot)$ implicitly through moment functions. Either we will be solving a set of moment conditions:

$$E_\theta \psi(Z, \beta) = 0, \tag{2}$$

where $\psi$ is a given function and there is a unique solution for all $\theta \in \Theta$; or we shall be minimizing a moment function:

$$\beta = \arg\min_t E_\theta[\rho(Z, t)], \tag{3}$$

where $\rho$ is a given function and there is a unique solution for all $\theta \in \Theta$. Note that the dimension of $\psi$ in (2) equals the dimension of $\beta$, so that no restrictions are being imposed on the $P_\theta$ distribution.

There is a potential pitfall in using the Dirichlet prior with large $J$ and all of the $b_j$ bounded away from zero. To see this, let $\phi$ denote the probability that $Z$ is in some set $B$: $\phi = \sum_{j:a_j \in B} \theta_j$. Then the posterior distribution for $\phi$ is a beta distribution with

$$E(\phi \mid d) = \sum_{j:a_j \in B} (n_j + b_j) \Big/ \sum_{j=1}^J (n_j + b_j)$$

$$\mathrm{Var}(\phi \mid d) = E(\phi \mid d)[1 - E(\phi \mid d)]/(1 + \sum_{j=1}^J (n_j + b_j)).$$

Suppose that $b_j = \epsilon > 0$ for all $j$, and consider increasing the number of support points while keeping the data $d$ fixed. Let the fraction of support points in $B$ approach a limit $r$: $\frac{1}{J}\sum_{j=1}^J 1(a_j \in B) \to r$ as $J \to \infty$. Then $E(\phi \mid d) \to r$, $\mathrm{Var}(\phi \mid d) \to 0$, and the posterior

distribution of $\phi$ becomes concentrated on $r$, regardless of the data. This argument covers a flat prior for $\theta$ ($b_j \equiv 1$), suggesting that a flat prior distribution does not capture a lack of prior information very well when $J$ is large. Therefore, we focus on the improper prior distribution with all the $b_j \to 0$.

The algorithm for evaluation of $\beta = g(\theta)$ defined through moment functions takes a particularly simple form for the limiting posterior distribution that results from letting all the $b_j \to 0$ in (1). Then the $\theta_j$ corresponding to the support points $a_j$ not observed in the sample are all zero with posterior probability one. Let $\{V_i\}_{i=1}^n$ be independently distributed according to a standard exponential distribution (i.e., the gamma distribution $\mathcal{G}(1,1)$). Then for a given function $\lambda(\cdot)$, the posterior distribution of $E_\theta[\lambda(Z)]$ is the same as the distribution of $\sum_{i=1}^n \lambda(z_i)V_i / \sum_{i=1}^n V_i$ since

$$\sum_{i=1}^{n} \lambda(z_i)V_i \Big/ \sum_{i=1}^{n} V_i = \sum_{j:n_j>0} \lambda(a_j)U_j \Big/ \sum_{j:n_j>0} U_j,$$

where $U_j = \sum_{i:z_i=a_j} V_i \sim \mathcal{G}(n_j, 1)$, using the fact that a sum of independent exponential random variables has a gamma distribution. So to simulate the posterior distribution of $\beta$ based on (2), we draw sets of i.i.d. exponential random variables $\{V_i^{(l)}\}_{i=1}^n$ and solve

$$\sum_{i=1}^{n} \psi(z_i, \beta^{(l)})V_i^{(l)} = 0, \tag{4}$$

and for $\beta$ based on (3) we solve

$$\beta^{(l)} = \arg\min_t \sum_{i=1}^{n} \rho(z_i, t)V_i^{(l)}. \tag{5}$$

Repeating this for $l = 1, \ldots, L$ gives us $L$ independent draws from the posterior distribution of $\beta$. Note that there is no need to divide by the sum of the exponential draws. Rubin (1981) developed this simulation algorithm, and it has been applied by Lancaster (1994) in the analysis of choice-based samples.[3]

The improper prior distribution for $\theta$ does not imply a unique prior distribution for the parameter of interest.[4] So in order to measure the informativeness of the prior distribution, we calculate the expected posterior distribution given a small number $m$ of observations, where we take the expectation over the empirical distribution. Let $F_n$ denote the empirical distribution of our sample: $F_n(B) = \frac{1}{n}\sum_{i=1}^{n} 1(z_i \in B)$. Let $\pi_m^\beta(\cdot \mid \{t_i\}_{i=1}^m)$ denote the posterior distribution for $\beta$ based on the $m$ observations $Z_i = t_i$ (and assume for a moment that this posterior distribution is proper). The expected posterior distribution for $\beta$ based on a random sample (with replacement) of size $m$ from $F_n$ is given by $\bar{\pi}_m^\beta(\cdot) = \int \pi_m^\beta(\cdot \mid \{t_i\}_{i=1}^m) \prod_{i=1}^m dF_n(t_i)$. In order to allow for the possibility of an improper posterior distribution, we modify this formula as follows:

$$\bar{\pi}_m^\beta(\cdot) = \tag{6}$$
$$\int \pi_m^\beta(\cdot \mid \{t_i\}_{i=1}^m) 1(\{t_i\}_{i=1}^m \in C_m) \prod_{i=1}^m dF_n(t_i) \Big/ \int 1(\{t_i\}_{i=1}^m \in C_m) \prod_{i=1}^m dF_n(t_i),$$

where the set $C_m$ consists of the points $\{t_i\}_{i=1}^m$ such that $\pi_m^\beta(\cdot \mid \{t_i\}_{i=1}^m)$ is a proper distribution. (We shall choose $m$ large enough so that the probability in the denominator of (6) is nonzero.) We can simulate this distribution as follows. In order to obtain the draw $\beta^{(l)}$, draw random samples of size $m$ from the empirical distribution $F_n$ until we obtain a sample $\{t_i^{(l)}\}_{i=1}^m$ in $C_m$; then draw a single $\beta^{(l)}$ from the posterior distribution $\pi_m^\beta(\cdot \mid \{t_i^{(l)}\}_{i=1}^m)$. Repeating this process $L$ times gives a random sample $\{\beta^{(l)}\}_{l=1}^L$ from $\bar{\pi}_m^\beta$.

For an example of $\bar{\pi}_m^\beta$ in a simple parametric case, consider sampling from a bivariate normal distribution with unknown mean $\theta = (\theta_1, \theta_2)$ and known covariance matrix $\Sigma$. The standard diffuse prior has density $p(\theta) \propto 1$. It can be obtained as the limit as $s \to \infty$ of a $\mathcal{N}(0, sD)$ prior, where $D$ may be any $2 \times 2$ positive-definite matrix. Suppose that the parameter of interest is $\beta = \theta_1/\theta_2$. Then the implied prior for $\beta$ is the distribution of $W_1/W_2$, where $(W_1, W_2) \sim \mathcal{N}(0, D)$. So the implied prior for $\beta$ depends upon the choice of $D$. Given a sample of size $m$ with mean $\bar{t}_m$, the posterior distribution for $\theta$ is $\mathcal{N}(\bar{t}_m, \frac{1}{m}\Sigma)$, which does not depend upon $D$. In order to sample from the $\bar{\pi}_m^\beta$ distribution, draw a

random sample $\{t_i\}_{i=1}^m$ from $F_n$, draw $\theta$ from $\mathcal{N}(\bar{t}_m, \frac{1}{m}\Sigma)$, and then set $\beta = \theta_1/\theta_2$. If $m = 1$, this reduces to drawing $\theta$ from the convolution of $F_n$ and $\mathcal{N}(0, \Sigma)$, and setting $\beta = \theta_1/\theta_2$.

Our framework of combining a multinomial likelihood with a Dirichlet prior makes it difficult to impose restrictions on the family of probability distributions $\{P_\theta, \theta \in \Theta\}$. That is why we are not treating the generalized method-of-moments case where the dimension of the moment function $\psi$ in (2) may exceed the dimension of the parameter $\beta$ (as in Hansen (1982))[5]. It is also difficult to impose smoothness restrictions. The improper prior distribution for $\theta$ results in a posterior distribution that assigns zero probability to the support points $a_j$ that are not observed in the sample. More generally, consider a proper prior in which the sum of the prior parameters $\sum_{j=1}^J b_j$ is small relative to the sample size $n$, so that the prior does not dominate the sample. If $J >> n$, then most of the $b_j$ will be very small. Consider two support points, $a_j$ and $a_k$, with $b_j$ and $b_k$ very small. If $a_j$ is observed in the sample and $a_k$ is not, then the ratio $P(Z = a_j \mid d)/P(Z = a_k \mid d)$ will be large even if $a_j$ and $a_k$ are close together. So our framework is not well suited for imposing smoothness on the $\{P_\theta, \theta \in \Theta\}$ family. See Diaconis and Freedman (1986) for more discussion of this issue.

## 3. INSTRUMENTAL VARIABLES

We shall use a very simple model with a constant, additive treatment effect, linear in years of schooling. The potential outcome with treatment level $s$ is

$$Y_s = Y_0 + \gamma s,$$

where $Y_0$ is the potential outcome with treatment level 0, and $\gamma$ is the unknown return to schooling. The actual treatment level is $X$, which gives an actual outcome $Y$ of

$$Y = Y_0 + \gamma X.$$

The potential outcome $Y_0$ has a linear predictor $\alpha' R$ based on the observed regressors $R$ (with $\alpha \equiv \arg\min_t E_\theta(Y_0 - t'R)^2$); then defining the disturbance $U = Y_0 - \alpha' R$ gives the orthogonal decomposition

$$Y_0 = \alpha' R + U, \quad E_\theta(RU) = 0.$$

Here the first component of $R$ is a constant identically equal to one, so that $E_\theta(U) = 0$. The instrumental variable $W$ satisfies $E_\theta(WU) = 0$ and $\mathrm{Cov}_\theta(W, X) \neq 0$.

Let $Z = (Y, X, R, W)$ and $\beta' = (\alpha', \gamma)$. Then $\beta$ satisfies the moment condition $E_\theta \psi(Z, \beta) = 0$ with

$$\psi(Z, \beta) = (Y - \alpha' R - \gamma X) \begin{pmatrix} R \\ W \end{pmatrix}.$$

We shall use the improper Dirichlet prior (with all the $b_j \to 0$ in (1)), and the posterior distribution of $\beta$ can be simulated as in (4).

Our data is a subset of the data used by Angrist and Krueger (1991) containing males born in either the first or fourth quarters between 1930 and 1939. The sample size is $n = 162,515$. The outcome variable $Y$ is the log of weekly earnings in 1979. The treatment $X$ is years of schooling completed, and the instrumental variable $W$ is an indicator equal to one if the individual was born in the fourth quarter and equal to zero otherwise. The regressor $R$ is simply a constant.[6]

In order to evaluate the information content of the prior distribution for the parameter of interest ($\gamma$), we shall calculate the expected posterior distribution $\bar{\pi}_m^\gamma$ as in (6), with $m = 3$ and $m = 10$ observations. We shall compare these expected posteriors with the actual posterior distribution based on the full sample with $n = 162,515$ observations. Here are some of the quantiles for the $\gamma$ distributions:

| quantile: | .025 | .05 | .25 | .50 | .75 | .95 | .975 |
|---|---|---|---|---|---|---|---|
| $\bar{\pi}_3^\gamma$: | -1.82 | -.82 | -.07 | .07 | .22 | 1.02 | 1.89 |
| $\bar{\pi}_{10}^\gamma$: | -2.43 | -1.02 | -.09 | .07 | .23 | 1.22 | 2.51 |
| $\pi_n^\gamma(\cdot \mid d)$: | .047 | .054 | .075 | .089 | .104 | .124 | .132 |

It appears that the prior distribution is reasonably uninformative for $\gamma$, so that the posterior distribution is mainly reflecting the sample information.[7]

The instrumental-variables estimate $\hat{\gamma}$ (i.e., the solution to $\sum_{i=1}^{n} \psi(z_i, \hat{\beta}) = 0$, where $\hat{\beta}' = (\hat{\alpha}', \hat{\gamma})$) is .089. An asymptotic approximation to its sampling distribution (allowing for heteroskedasticity of unknown form) gives a normal distribution with mean $\gamma$ and standard deviation .021. A normal distribution with mean .089 and standard deviation .021 provides a good approximation to our posterior distribution. But if the objective is a posterior distribution for $\gamma$, then our procedure is more direct than having to first approximate a sampling distribution and then argue that the sampling distribution can be used to approximate a posterior distribution. See Sims and Uhlig (1991) for a discussion of the distinction between a sampling distribution and a posterior distribution, in a case where the two need not coincide even asymptotically.

Suppose that the individual knows $Y_0 = y_0$ and wants to value the potential earnings distribution for various levels of $s$, as a first step in choosing the optimal level of education. The standard way to do this based on the microeconomics of decision making under uncertainty is to specify a utility function and evaluate the expected utility corresponding to various values of $s$. Suppose that utility is the following function of earnings and education:

$$u(earn, s) = w[earn/q(s)] - c(s),$$

where $w(\cdot)$ has the constant relative risk aversion form $w(t) = t^{(1-A)}/(1-A)$ for $A \neq 1$ and $w(t) = \log(t)$ for $A = 1$. The individual chooses $s$ to

$$\max_s E[u(e^{(y_0+\gamma s)}, s) \mid d].$$

We can define a certainty equivalent rate of return, which depends on $s$ but not on $y_0$:

$$E(e^{\gamma s(1-A)} \mid d) = e^{\gamma_{ce}(s)s(1-A)} \quad \Rightarrow$$

$$\gamma_{ce}(s) = \log[E(e^{\gamma s(1-A)} \mid d)]/(s(1-A)) \qquad (A \neq 1),$$

with $\gamma_{ce}(s) = E(\gamma \mid d)$ if $A = 1$. Then the optimal level of education can be obtained from

$$\max_s u(e^{(y_0+\gamma_{ce}(s)s)}, s).$$

To calculate $\gamma_{ce}(s)$ requires a distribution for $\gamma$, and we shall use the posterior distribution corresponding to the improper Dirichlet prior and the Angrist-Krueger data.

We find the following certainty equivalent values corresponding to $s = 8$, 12, 16 and various values for the coefficient ($A$) of relative risk aversion:

Certainty Equivalent $\gamma_{ce}$

| | A | | | | | |
|---|---|---|---|---|---|---|
| $s$ | 0 | 1 | 5 | 10 | 15 | 20 |
| 8 | .091 | .089 | .082 | .073 | .065 | .057 |
| 12 | .092 | .089 | .078 | .066 | .055 | .047 |
| 16 | .093 | .089 | .075 | .059 | .047 | .040 |

This calculation clearly requires a posterior distribution for $\gamma$, not a sampling distribution for $\hat{\gamma}$. More generally, a posterior distribution is called for in order to include parameter uncertainty in the decision making formulation; see, for example, Rossi, McCulloch, and Allenby (1994), Kandel and Stambaugh (1995), and Barberis (1996).

## 4. QUANTILE REGRESSION

Let $Z = (X, Y)$, where $Y$ is scalar and $X$ is $K \times 1$. We can define a linear predictor corresponding to the $\tau^{\text{th}}$ quantile as follows: $E^*_\theta(Y \mid X = x) = \beta' x$, where

$$\beta = \arg\min_t E_\theta[c_\tau(Y - t'X)]$$

$$c_\tau(t) = |t|[(1-\tau)1(t < 0) + \tau 1(t \geq 0)].$$

($\beta$ in general depends upon $\tau$, but this should be clear from the context.) If $\tau = .5$, then this reduces to minimizing the mean absolute error: $\min_t E_\theta(|Y - t'X|)$. By weighting the

absolute error differently for positive and negative values, the "check" function $c_\tau$ extends this notion of linear predictor to other quantiles. The role of the check function in quantile regression was developed by Koenker and Bassett (1978, 1982).

Our simulation procedure produces independent draws $\{\beta^{(l)}\}_{l=1}^{L}$ from the posterior distribution of $\beta$. To obtain $\beta^{(l)}$, first take i.i.d. draws $\{V_i^{(l)}\}_{i=1}^{n}$ from a standard exponential distribution ($\mathcal{G}(1,1)$). Then solve

$$\beta^{(l)} = \arg\min_t \sum_{i=1}^{n} V_i^{(l)} c_\tau(y_i - t'x_i)$$

(where the observed value of $Z_i$ is $z_i = (x_i, y_i)$). The computations are simplified by exploiting the fact that $rc_\tau(t) = c_\tau(rt)$ if $r \geq 0$. So define $Y_i^{(l)} = V_i^{(l)} y_i$ and $X_i^{(l)} = V_i^{(l)} x_i$. Then

$$\beta^{(l)} = \arg\min_t \sum_{i=1}^{n} c_\tau(Y_i^{(l)} - t'X_i^{(l)}).$$

This is a linear programming problem, and we use the Barrodale-Roberts (1973) modification of the standard simplex algorithm.

Our application is based on "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," by Meyer, Viscusi, and Durbin (1995). The authors (MVD) obtained data for two states, Kentucky and Michigan, on a random sample of indemnity claims. We shall focus on Kentucky. The claims were filed by workers seeking compensation for work-related injury or illness. MVD concentrate on temporary total disability claims. Such a claim is filed when the person is unable to work but is expected to recover fully and return to work. The data include date injured, duration of temporary total benefits, total medical costs, previous wage, weekly benefit amount, type of injury (body part affected and the type of damage), age, sex, marital status, and an industry code.

The amount of the weekly benefit is based on a schedule that determines the benefit as a function of previous earnings. The schedule has a ceiling, with earnings levels above a

threshold corresponding to the same weekly benefit. Kentucky raised the maximum benefit from \$131 to \$217 per week on July 15, 1980.

MVD work with claims that have injury dates during the year before or the year after the change in the benefit schedule. They also limit the sample to a high earnings group and a low earnings group. The weekly benefit amount for the high earnings group was affected by the increase in the benefit ceiling, whereas the benefit amount for the low earnings group was not affected. So the low earnings group can provide a control for period effects. The basic specification in MVD is

$$E_\theta(Y \mid X = x) = \beta_1 + \beta_2 x_2 \cdot x_3 + \beta_3 x_2 + \beta_4 x_3. \tag{7}$$

($x_1 \equiv 1$ denotes a constant.) Here $Y$ = log of duration, with duration measured by weeks of temporary total benefits paid; $x_2 = 1$ if injured after the benefit increase, $x_2 = 0$ otherwise; $x_3 = 1$ if high earnings group, $x_3 = 0$ otherwise. The key coefficient is $\beta_2$, measuring the effect of the benefit increase on time out of work, with controls for period and for the earnings group:

$$\begin{aligned}\beta_2 = &[E_\theta(Y \mid x_2 = 1, x_3 = 1) - E_\theta(Y \mid x_2 = 0, x_3 = 1)] \\ &- [E_\theta(Y \mid x_2 = 1, x_3 = 0) - E_\theta(Y \mid x_2 = 0, x_3 = 0)].\end{aligned}$$

An appealing aspect of the MVD analysis is that it is plausible to regard the injury date, and hence the benefit schedule, as if it were randomly assigned.

To account for possible changes in the composition of the sample after the benefit increase, MVD also include regression controls for attributes of the individual, the job, and the injury—sixteen regressors in addition to the four in (7). The last column of Table 1 presents least-squares estimates (and conventional standard errors) corresponding to Table 6 in MVD. The first five columns of Table 1 present estimates of the linear predictor coefficients corresponding to the .10, .25, .50, .75, and .90 quantiles. These estimates are based on the simulation procedure described above. The point estimates are

posterior medians and the "standard errors" in parentheses are constructed so that the point estimate plus or minus 1.96 standard errors gives an interval with a .95 posterior probability. The key coefficients (corresponding to $\beta_2$ in (7)) are in the second row. The effect of the benefit increase is fairly constant across the quantiles, suggesting a location model in which the distribution of log duration shifts rigidly in response to the benefit increase.

Table 2 presents results using duration out of work (in weeks) instead of its logarithm. Now the estimates show a substantial increase as we go from low to high quantiles, suggesting that the effect of the benefit increase is concentrated on the upper half of the duration distribution. The estimated effect on the median of the distribution is .87 weeks, with a standard error of .23. In contrast, the least-squares estimate of the effect on the mean of the distribution is quite imprecise, with a point estimate of 1.66 and a standard error of 1.04.

The histogram of the draws from the posterior distribution of $\beta_2$ is shown in Figure 1 for $\tau = .5$, using duration in weeks. The posterior mean is .87, and the posterior standard deviation is .23. So assuming the posterior distribution is normal and using $.87 \pm 1.96 \times .23$ gives a probability interval close to the one we constructed without assuming normality.

We shall examine the influence of the prior distribution by calculating the expected posterior distribution $\bar{\pi}_m^\beta$ as in (6), for $m = 21$ observations, and comparing this distribution with the posterior distribution $\pi_n^\beta(\cdot \mid d)$ based on the full sample with $n = 5349$ observations. Here are some of the quantiles of the $\beta_2$ distributions for $\tau = .5$, using duration in weeks:

| quantile: | .025 | .05 | .25 | .50 | .75 | .95 | .975 |
|---|---|---|---|---|---|---|---|
| $\bar{\pi}_{21}^{\beta_2}$: | -290 | -157 | -20.4 | 1.01 | 24.3 | 184 | 323 |
| $\pi_n^{\beta_2}(\cdot \mid d)$: | .41 | .49 | .71 | .87 | 1.03 | 1.25 | 1.32 |

The prior distribution is dominated by the sample information.

Now consider dropping all the predictor variables except for the four that appear in (7): 1, $x_2 \cdot x_3$, $x_2$, $x_3$. We shall compare the expected posterior distribution for $m = 5$ observations with the posterior distribution based on the full sample. Here are quantiles of these distributions for $\beta_2$ with $\tau = .5$, using duration in weeks:

| quantile: | .025 | .05 | .25 | .50 | .75 | .95 | .975 |
|---|---|---|---|---|---|---|---|
| $\bar{\pi}_5^{\beta_2}$: | -121 | -36 | -6 | 1 | 9 | 59 | 110 |
| $\pi_n^{\beta_2}(\cdot \mid d)$: | 0 | 0 | 1 | 1 | 2 | 2 | 2 |

The posterior histogram for $\beta_2$ is in Figure 2. It is concentrated on just four points: -1, 0, 1, and 2 weeks, with posterior probabilities of .01, .14, .55, and .30. This reflects the discreteness of the duration distribution. The upper tail of that distribution is somewhat continuous, but 56% of the distribution is concentrated on the integers from 0 to 4 weeks. The $(.5, .75, .9, .95, .975)$ quantiles are $(4, 8, 15, 25, 49)$ weeks. Including the long list of predictor variables smoothes out this discreteness, in the sense of producing a residual distribution (for $Y - \beta' X$) that is much closer to being continuous.

Here are the quantiles of the $\beta_2$ distributions for $\tau = .9$, using just the four regressors in (7) and duration in weeks:

| quantile: | .025 | .05 | .25 | .50 | .75 | .95 | .975 |
|---|---|---|---|---|---|---|---|
| $\bar{\pi}_5^{\beta_2}$: | -145 | -41 | -7 | 1 | 10 | 72 | 124 |
| $\pi_n^{\beta_2}(\cdot \mid d)$: | 2 | 3 | 5 | 7 | 8 | 11 | 12 |

The posterior histogram for $\beta_2$ is in Figure 3. This is closer to a normal distribution, corresponding to the continuity in the upper tail of the duration distribution.

The standard asymptotic distribution theory for quantile regression requires that the distribution of $Y - \beta' x$ (conditional on $\theta$ and on $X = x$) should be absolutely continuous with a positive density in a neighborhood of zero. This theory may be a reasonable guide when we include the long list of predictor variables. It is certainly not a reasonable guide when we just use the two indicator variables and their interaction. In contrast,

our posterior distributions provide straightforward inferences that do not rely upon the approximate normality of a sampling distribution.

Efron's (1979) bootstrap method suggests an alternative approach to inference. It provides an approximate sampling distribution for an estimator by treating the empirical distribution of the sample as if it were the population distribution. We can obtain a draw $\hat{\beta}^{(l)}$ from the bootstrap distribution for $\hat{\beta}$ as follows. Draw a random sample of size $n$, $\{(\tilde{X}_i^{(l)}, \tilde{Y}_i^{(l)})\}_{i=1}^n$, from the empirical distribution $F_n$ of the sample (with replacement), and then solve

$$\hat{\beta}^{(l)} = \arg\min_t \sum_{i=1}^n c_\tau(\tilde{Y}_i^{(l)} - t'\tilde{X}_i^{(l)}).$$

Rubin (1981) argued that the bootstrap distribution will tend to be similar to the posterior distribution corresponding to an improper Dirichlet prior, and he labeled the simulation procedure in (4) and (5) the "Bayesian bootstrap." This similarity is apparent in our application. Using the long list of regressors (and duration in weeks), the histogram of the draws from the bootstrap distribution for $\hat{\beta}_2$ with $\tau = .5$ is very similar to Figure 1. The mean is .87 weeks with a standard deviation of .23, matching the posterior distribution. With the short list of regressors, the bootstrap distributions for $\tau = .5$ and .9 closely resemble the posterior distributions in Figures 2 and 3. With $\tau = .5$, the bootstrap distribution is concentrated on -1, 0, 1, and 2 weeks with probabilities of .01, .14, .57, and .28. With $\tau = .9$, the quantiles of the bootstrap distribution are

| quantile: | .025 | .05 | .25 | .50 | .75 | .95 | .975 |
|---|---|---|---|---|---|---|---|
| Bootstrap: | 2 | 3 | 5 | 7 | 8 | 11 | 12 |

The only available justification for the Efron bootstrap in quantile regression relies on an asymptotic distribution theory that requires $E_\theta[c_\tau(Y - t'X)]$ to have a nonsingular second derivative at $t = \beta$—see Hahn (1995). This is not a plausible condition when we use the short list of regressors. Nevertheless, we can interpret the Efron bootstrap as providing a close approximation to the posterior distribution based on the improper Dirichlet prior.

## 5. CONCLUSION

The Bayesian approach to inference provides an attractive conceptual framework due to its connection with optimization concepts in decision theory and its lack of reliance on large-sample approximations. In practice, its use has been limited by the requirement of a fully specified parametric model since many econometric models are only partly specified. In this paper we have presented two applications of a less parametric Bayes approach, due to Ferguson (1973, 1974) and Rubin (1981). In the first application, the decision-theoretic nature of the underlying question forces the use of posterior distributions rather than sampling distributions. In the second application, the assumptions underlying the asymptotic normality of the sampling distributions are clearly violated, but inference based on posterior distributions is straightforward.

We hope that future work will extend this approach to allow for restrictions on the $\{P_\theta, \theta \in \Theta\}$ family of distributions. The restrictions could arise from having more moment conditions than parameters ($\dim(\psi) > \dim(\beta)$ in (2)), or there could be smoothness restrictions. There has been recent progress in developing related approaches in mixture models, density estimation, and binary response models that might be useful in this respect; see, for example, Doss (1994), Kong, Liu, and Wong (1994), Escobar and West (1995), and Newton, Czado, and Chappell (1996).

## FOOTNOTES

[1] The authors thank David Cox, Jinyong Hahn, and Neil Shephard for helpful comments, and thank Alan Krueger and Bruce Meyer for making their data available to us. The National Science Foundation provided financial support.

[2] See Wilks (1962) and Ferguson (1973).

[3] The simulation procedure in Rubin (1981) is based on the $V_i/\sum_{i=1}^{n} V_i$ having an alternative representation as the gaps between the order statistics formed from a random sample of size $n-1$ from a uniform $(0,1)$ distribution.

[4] Consider letting $S \equiv \sum_{j=1}^{J} b_j \to 0$ with $(b_1/S, \ldots, b_J/S)$ held fixed at $\zeta$, where $\zeta$ is some point in $\Theta$ (the unit simplex in $\mathcal{R}^J$). Let $\mathcal{Z} = \{a_1, \ldots, a_J\}$ denote the sample space. The results of Sethuraman and Tiwari (1982) and Sethuraman (1994), specialized to a finite sample space, show that there is a limiting prior distribution on $\Theta$: if $\theta$ is a draw from that distribution, then $P_\theta$ is a point mass at $z \in \mathcal{Z}$, where $z$ is a draw from $P_\zeta$. Then the moment condition (2) defining $\beta$ becomes $\psi(z,\beta) = 0$—see Florens and Rolin (1994). So the implied prior distribution for $\beta$ depends upon $\zeta$; i.e., it depends upon the precise way in which the $b_j \to 0$. Furthermore, the condition $\psi(z,\beta) = 0$ may not uniquely define $\beta$ even given $\zeta$. For example, consider the linear predictor problem where $Z = (X,Y)$ and $\psi(z,\beta) = x(y - \beta' x)$. Then $\psi(z,\beta) = 0$ does not have a unique solution when $X$ includes more than one regressor. A similar lack of uniqueness arises in our instrumental variables application and in our quantile regression application.

[5] See Chamberlain and Imbens (1995) for an approach to this over-identified case.

[6] The results are similar when $R$ consists of 509 dummy variables obtained from interacting state-of-birth dummy variables with year-of-birth dummy variables—see footnote 7. (There are fifty states plus the District of Columbia and ten years, 1930–39; there are no observations for Alaska in 1931.)

[7] Note that $\bar{\pi}_{10}^{\gamma}$ is more dispersed than $\bar{\pi}_{3}^{\gamma}$; $\bar{\pi}_{100}^{\gamma}$ is similar to $\bar{\pi}_{10}^{\gamma}$ when $R = 1$. When $R$

consists of interactions of state-of-birth and year-of-birth dummy variables, the quantiles for the $\gamma$ distributions are

| quantile: | .025 | .05 | .25 | .50 | .75 | .95 | .975 |
|---|---|---|---|---|---|---|---|
| $\bar{\pi}^{\gamma}_{100}$: | -1.91 | -.99 | -.11 | .07 | .25 | 1.19 | 2.35 |
| $\pi^{\gamma}_{n}(\cdot \mid d)$: | .056 | .062 | .084 | .097 | .112 | .132 | .141 |

(In simulating the expected posterior distribution $\bar{\pi}^{\gamma}_{100}$ when $R$ consists of 509 dummy variables, we take random samples of size $m = 100$ from the empirical distribution $F_n$; 84 of the 509 state-year cells are nonempty on average for the samples that give proper posterior distributions for $\gamma$.)

## REFERENCES


Angrist, J., and Krueger, A. (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106, 979–1014.

Barberis, N. (1996): "Investing for the Long Run when Returns are Predictable," in *Essays on Financial Economics*, Ph.D. dissertation, Harvard University, chap. 1.

Barrodale, I., and Roberts, F. (1973): "An Improved Algorithm for Discrete $l_1$ Linear Approximation," *SIAM Journal of Numerical Analysis*, 10, 839–848.

Chamberlain, G., and Imbens, G. (1995): "Semiparametric Applications of Bayesian Inference," Harvard Institute of Economic Research, Discussion Paper No. 1716.

Diaconis, P., and Freedman, D. (1986): "On the Consistency of Bayes Estimates" (with discussion), *The Annals of Statistics*, 14, 1–67.

Doss, H. (1994): "Bayesian Nonparametric Estimation for Incomplete Data Via Successive Substitution Sampling," *The Annals of Statistics*, 22, 1763–1786.

Efron, B. (1979): "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

Escobar, M., and West, M. (1995): "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. (1973): "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230.

Ferguson, T. (1974): "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615–629.

Florens, J.-P., and Rolin, J.-M. (1994): "Bayes, Bootstrap, Moments," unpublished manuscrip Université des Sciences Sociales de Toulouse.

Hahn, J. (1995): "Bootstrapping Quantile Regression Estimators," *Econometric Theory*, 11, 105–121.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Kandel, S., and Stambaugh, R. (1995): "On the Predictability of Stock Returns: An Asset-Allocation Perspective," National Bureau of Economic Research, Working Paper No. 4997.

Koenker, R., and Bassett, G. (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R., and Bassett, G. (1982): "Robust Tests for Heteroscedasticity Based on Regression Quantiles," *Econometrica*, 50, 43–61.

Kong, A., Liu, J., and Wong, W. (1994): "Sequential Imputations and Bayesian Missing Data Problems," *Journal of the American Statistical Association*, 89, 278–288.

Lancaster, T. (1994): "Bayes WESML: Posterior Inference from Choice-Based Samples," unpublished manuscript, Brown University.

Meyer, B., Viscusi, W. K., and Durbin, D. (1995): "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review*, 85, 322–340.

Newton, M., Czado, C., and Chappell, R. (1996): "Bayesian Inference for Semiparametric Binary Regression," *Journal of the American Statistical Association*, 91, 142–153.

Rossi, P., McCulloch, R., and Allenby, G. (1994): "Hierarchical Modelling of Consumer Heterogeneity: An Application to Target Marketing," unpublished manuscript, University of Chicago.

Rubin, D. (1981): "The Bayesian Bootstrap," *The Annals of Statistics*, 9, 130–134.

Sethuraman, J. (1994): "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

Sethuraman, J., and Tiwari, R. (1982): "Convergence of Dirichlet Measures and the Interpretation of Their Parameter," in *Statistical Decision Theory and Related Topics III*, eds. S. Gupta and J. Berger, New York: Academic Press, pp. 305–315.

Sims, C., and Uhlig, H. (1991): "Understanding Unit Rooters: A Helicopter Tour," *Econometrica*, 59, 1991, 1591–1599.

Wilks, S. (1962): *Mathematical Statistics*, New York: Wiley.

TABLE 1

Quantile Regression Coefficients for Log of Duration, Kentucky
High and Low Earnings Groups Pooled

| | Quantile | | | | | |
|---|---|---|---|---|---|---|
| Variables | .10 | .25 | .50 | .75 | .90 | OLS |
| Intercept | -5.555 | -3.067 | -1.749 | -0.811 | -1.239 | -1.994 |
| | (0.817) | (0.497) | (0.403) | (0.490) | (0.692) | (0.410) |
| After increase | 0.136 | 0.141 | 0.164 | 0.170 | 0.137 | 0.145 |
| *High earnings group | (0.102) | (0.057) | (0.053) | (0.060) | (0.088) | (0.051) |
| After increase | -0.008 | -0.039 | -0.029 | 0.013 | 0.074 | 0.000 |
| | (0.073) | (0.042) | (0.034) | (0.040) | (0.057) | (0.033) |
| High earnings group | 1.755 | 0.525 | 0.024 | -0.792 | -3.191 | -0.696 |
| | (1.352) | (0.931) | (0.771) | (1.014) | (1.692) | (0.806) |

Note: The dependent variable is ln(.5 + duration). The sample size is 5349. The additional regressors are Ln(previous wage), Ln(previous wage)*High earnings group, Male, Married, Ln(age), Ln(total medical costs), Hospital stay indicator; *Industry indicators*: Manufacturing, Construction; *Injury type indicators*: Head, Neck, Upper extremities, Trunk, Low back, Lower extremities, Occupational diseases. The omitted industry is other industries, and the omitted injury is other injuries.

TABLE 2

Quantile Regression Coefficients for Duration, Kentucky
High and Low Earnings Groups Pooled

| | Quantile | | | | | |
|---|---|---|---|---|---|---|
| Variables | .10 | .25 | .50 | .75 | .90 | OLS |
| Intercept | -6.199 | -7.258 | -8.972 | -11.566 | -19.848 | -25.886 |
| | (1.157) | (1.441) | (1.779) | (3.310) | (7.254) | (8.412) |
| After increase | 0.229 | 0.302 | 0.873 | 1.351 | 2.661 | 1.665 |
| *High earnings group | (0.143) | (0.165) | (0.230) | (0.554) | (1.339) | (1.043) |
| After increase | -0.052 | -0.032 | -0.116 | 0.122 | 0.498 | 0.457 |
| | (0.085) | (0.097) | (0.138) | (0.289) | (0.629) | (0.674) |
| High earnings group | 0.051 | -0.356 | -1.655 | -11.541 | -56.802 | -41.783 |
| | (2.546) | (2.848) | (3.528) | (9.299) | (27.400) | (16.539) |

Note: The dependent variable is duration (in weeks). The sample size is 5349. The additional regressors are the same as in Table 1. The omitted industry is other industries, and the omitted injury is other injuries.

Figure 1. Posterior Histogram, q = .5 (long list for x)

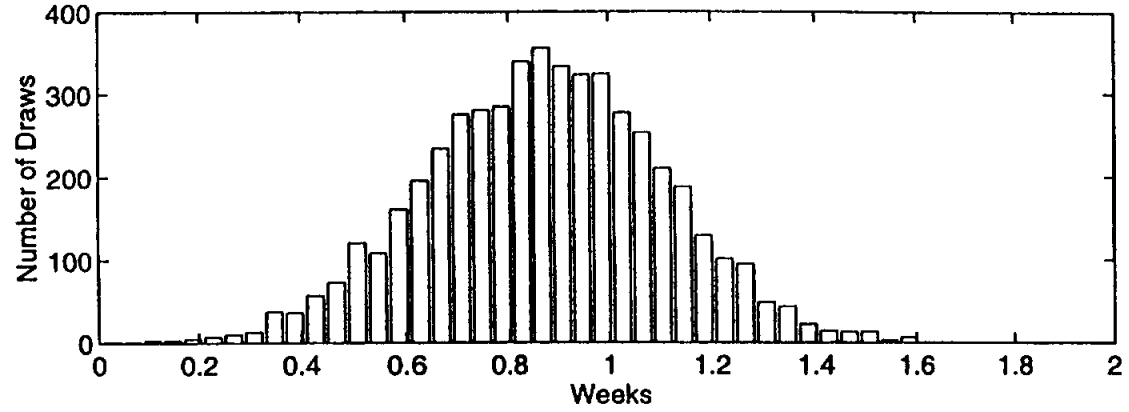


Figure 2. Posterior Histogram, q = .5 (short list for x)

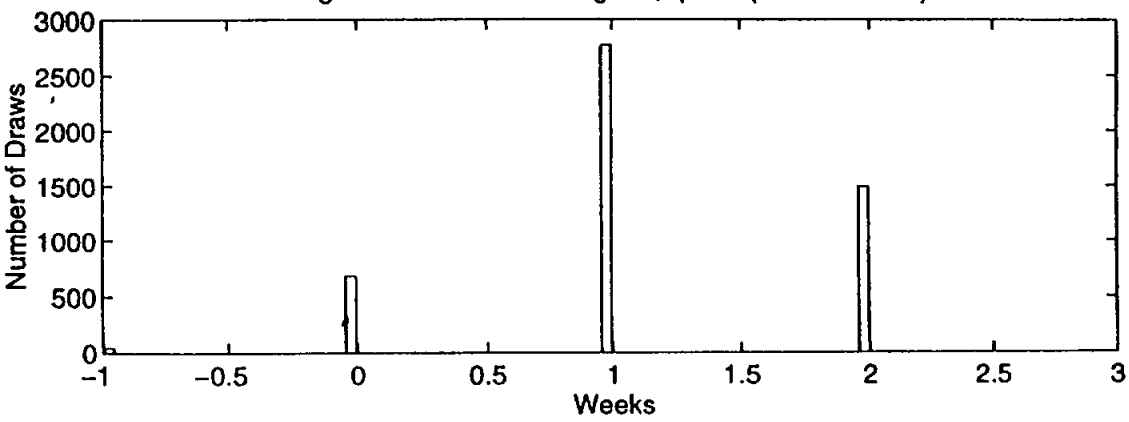

Figure 3. Posterior Histogram, q = .9 (short list for x)